DS 3001 Project Pre Analysis Plan

Kaitlyn Chou, Mackenzie Chen, Mrunal Kute

October 2025

1 Question

To reiterate, our research question is to determine whether general trends of COVID-19 are increasing, decreasing, or stagnant in the future, as well as whether the underlying health problems of the patient affect this trend.

2 Method Overview

The objective of this study is to determine whether the general trends of COVID-19 are increasing, decreasing, or stagnant in the future, as well as whether the underlying health problems of the patient affect this trend.

Prior to analysis, we plan to undergo a cleaning process to handle any missing values, outliers, or variables that are incorrectly formatted. To summarize from the previous work, in order to clean the data, we will first drop null values found in variables, such as "hospitalization", "death", or "ICU status". We hypothesize that these are the cases that were not followed up on. Additionally, we will further clean the data by converting variables in text format, like "status", into a numeric format so we can make the data a binary value (0 or 1). Additionally, for variables on "race/ethnicity", we can further use one-hot encoding to create separate binary columns for each ethnic group, such as *Black*, *White*,*Asian*, *Hispanic*, and so on. Each column would indicate whether a person identifies with that group (1) or not (0). For individuals who identify as mixed race, we would assign a 1 in multiple relevant columns. This approach allows the data to represent multiple identities without imposing a false numerical hierarchy or a distance between ethnicities. This will make it easier to report on the statistics in the data.

For analysis, we plan to use a decision tree model to analyze the COVID-19 data, specifically, a regression tree. A decision tree model was chosen for analysis for multiple reasons. For one, the model's ability to handle non-linear relationships is useful, as it can handle spikes in cases. Another reason why a decision tree model is the best model for our purposes is that it can automatically capture interactions between features in the model and is less sensitive

to outliers and external noise. Additionally, it is versatile and adaptable in its ability to incorporate multiple data types, such as numeric or categorical data, easily. Finally, tree ensembles, like RF or XGBoost are commonly used in real COVID prediction systems, proving their effectiveness in these types of predictive algorithms. A regression tree was chosen over a classification tree because the target variable represents a continuous numerical outcome, such as the projected rate or number of future COVID-19 cases, rather than a categorical classification.

We chose not to use the kNN model for our data because firstly, it does not model time well, and it is sensitive to scale and noise. Most importantly, it is slow for large datasets, which will not be helpful for our dataset, which is an extremely large dataset with plenty of information on COVID-19 cases. Additionally, the biggest drawback for the kNN model is that it has poor extrapolation; it cannot predict beyond the range of observed data, which is not helpful for our model because we want to use it to predict COVID-19 trends.

To train the model, the dataset will first be prepared by organizing the input features into distinct training and testing sets. Within the training phase, the model will mix and evaluate groups of data points to identify potential splits using the Sum of Squared Errors (SSE) criterion. We plan to use a regression tree to predict numeric outcomes. To construct the regression tree, a potential split point is selected for each predictor variable. For every possible split, the mean value of the outcome variable is calculated on both sides of the split. The model then computes the Sum of Squared Errors (SSE) for that split, representing the total squared deviation of each observation from its group mean. We plan to select the split that minimizes the SSE, therefore producing the most homogeneous groups in terms of their outcome values. This recursive process will continue down the tree, with each new split further reducing the overall SSE. Because the tree model will be a regressor tree (to predict numeric outcomes), stopping rules such as maximum tree depth or minimum node size will be established to prevent overfitting.

3 Model Validation and Justification

To validate our regression model, we will compare the predicted COVID-19 trend values to the actual observed outcomes in the test dataset. Specifically, we will use the Mean Squared Error (MSE) and the Coefficient of Determination (R^2) to assess how well the model captures variation in the data. The MSE measures the average squared difference between the predicted and actual values, where lower values indicate better predictive accuracy. The R^2 value represents the proportion of variance in the observed data that can be explained by the model, with values closer to one representing a stronger fit. Additionally, we will also plot a scatterplot of predicted versus actual values to visually assess the model's accuracy and identify potential unique patterns and/or errors.